

[WEDT.A] Klasyfikacja typów serwisów WWW na podstawie informacji o strukturze strony i tekstu

Michał Aniserowicz, Jakub Turek

Opis problemu

Zadanie polega na implementacji aplikacji, która dokonuje automatycznej klasyfikacji typów stron WWW na podstawie ich struktury. Analiza może obejmować źródło strony, konfigurację rozmieszczenia komponentów (layout), a także strukturę i znaczenie zamieszczonych na stronie treści.

Założenia

Projekt obejmuje implementację klasyfikatora następujących typów serwisów:

Blog rodzaj internetowego dziennika (pamiętnika), który zawiera odrębnie, chronologicznie uporządkowane wpisy. Przykład serwisu: <http://rafalstec.blox.pl/>.

Serwisy informacyjne portale zawierające najnowsze wiadomości z różnych dziedzin życia, takich jak polityka, finanse, technologie. Przykład serwisu: <http://onet.pl/>.

„Kwejki” serwisy społecznościowe oparte w głównej mierze na grafikach. Przykład serwisu: <http://kwejk.pl/>.

Sklepy internetowe portale umożliwiające zakup szerokiego asortymentu akcesoriów elektronicznych oraz komputerowych. Przykład serwisu: <http://wicompl/>.

Dane wejściowe aplikacji stanowić będzie adres witryny internetowej. Na wyjście wyprowadzona zostanie nazwa kategorii lub informacja, że serwis nie został zaklasyfikowany do żadnej z powyższych kategorii.

Technologia

Projekt zostanie zaimplementowany w języku Python, przy wykorzystaniu wersji drugiej (2.7.4) interpretera języka. Implementacja będzie testowana w środowiskach Windows oraz Unix (Ubuntu). Do implementacji zostaną wykorzystane standardowe moduły języka, między innymi:

re wyrażenia regularne,

htmllib parser języka HTML.

```

<div class="tooltip-title-container">
  <div class="tooltip-title-left-corner">
    <div class="tooltip-title">
      <p class="tooltip-title-h2">
        <a href="/obrazek/1763501/autor-gry-o-tron.html">
          Autor Gry o Tron?
        </a>
      </p>
    </div>
  <div class="tooltip-title-right-corner"></div>
  <div class="clr"></div>
</div>
</div>

```

Rysunek 1: Fragment kodu źródłowego witryny <http://kwejk.pl>.

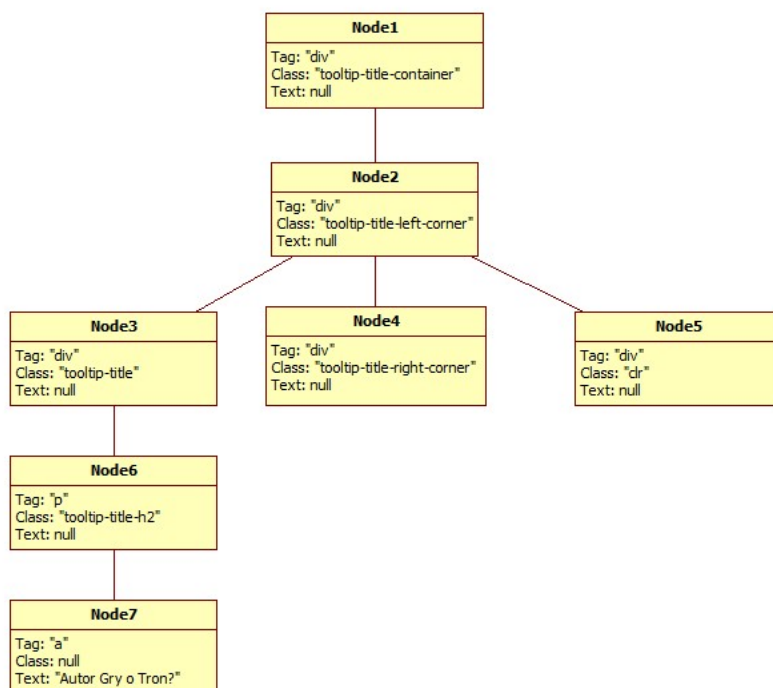
Struktura danych

Aplikacja umożliwiać będzie budowanie pełnego drzewa HTML. W korzeniu drzewa przechowywane będą następujące informacje:

- Typ napotkanego taga HTML, na przykład `<div>`, `<h1>`.
- Dodatkowe atrybuty taga powiązane z CSS - kaskadowymi arkuszami styli: identyfikator `id="_"`, klasa `class="_"` oraz styl elementu `style="_"`.
- Tekst zawarty pomiędzy tagiem otwierającym a zamykającym. Przykładowo dla kodu `<a>Odknośnik` jest to fraza „Odknośnik”.
- Inne atrybuty kontekstowe związane z poszczególnymi tagami:
 - dla obrazka (``) - jego rozmiar oraz źródło pochodzenia (lokalne - z domeny, którą analizujemy lub zewnętrzne - spoza niej),
 - dla nagłówków (`<h1>`, `<h2>`, itd.) - rozmiar czcionki.

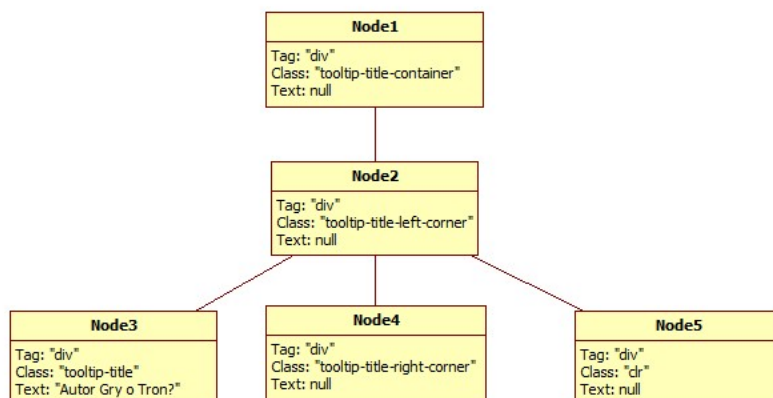
Ze względu na rozmiary oraz skomplikowanie struktury dla dużych portali, takich jak sklepy internetowe lub serwisy informacyjne, kod aplikacji będzie udostępniał różne możliwości redukcji złożoności drzewa:

- Zawężanie podzbioru tagów, dla których budowane jest drzewo. Tagi istotne dla struktury strony to, między innymi, `<div>`, `<td>`, `<article>`, `<h1>`, `<a>` oraz ``. Z punktu widzenia zadania, tagi niosące niewiele informacji służą głównie do formatowania tekstu, jak na przykład ``, ``, oraz osadzania skryptów - `<script>`.
- Ograniczanie stopnia zagnieżdżenia korzeni w drzewie:
 - pomijanie węzłów przekraczających dany, parametryzowalny, poziom zagnieżdżenia w strukturze,
 - sklepanie kilku następujących po sobie węzłów o zbliżonych wymiarach na stronie w jeden.



Rysunek 2: Pełne drzewo HTML dla kodu przedstawionego na listingu 1.

- Odfiltrowywanie elementów uznanych za nieistotne metodami heurystycznymi, przykładowo prosty filtr eliminujący reklamy bazując na klasach obiektów.

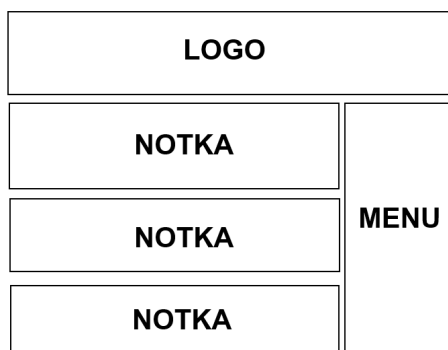


Rysunek 3: Drzewo HTML z rysunku 2 zredukowane do tagów div.

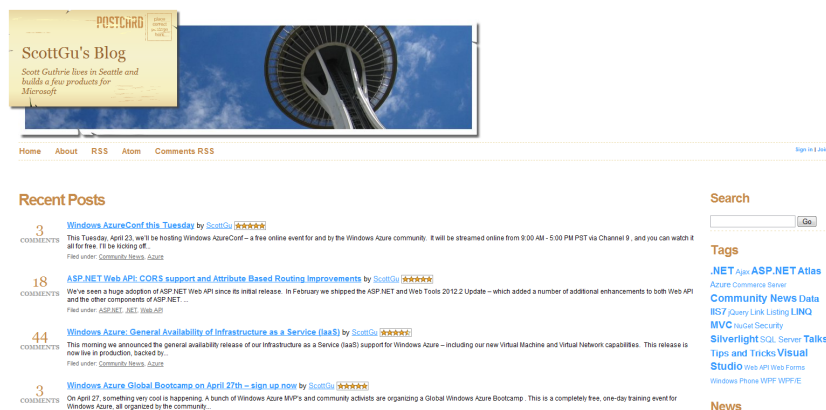
Algorytm

Blogi

Algorytm kategoryzacji blogów będzie rozpoznawał dzienniki internetowe w dwukolumnowym układzie. W szerszej kolumnie znajdują się notki, natomiast węższa kolumna to menu strony. Notki posiadają pewną stałą strukturę wewnętrzną. Na pojedynczy wpis składają się: tytuł, treść, data, informacja o autorze oraz link do komentarzy. Menu charakteryzuje się natomiast występowaniem dużej ilości podobnych odnośników jeden pod drugim (archiwum dla kolejnych miesięcy, kategorie wpisów, odnośniki do zaprzyjaźnionych blogów). Szablon bloga przedstawia rysunek 4. Przykłady stron o takiej strukturze przedstawiono na rysunkach 5 oraz 6.



Rysunek 4: Szablon, na którym oparta jest większość współczesnych blogów.



Rysunek 5: Przykład dwukolumnowej struktury bloga.

Algorytm rozpoznawania blogów będzie działał według następującego schematu:

1. Odnalezienie w drzewie powtarzającego się układu węzłów, które reprezentują notkę. Cechy charakterystyczne takiego układu to:



Rysunek 6: Przykład dwukolumnowej struktury bloga.

- Duża ilość tekstu zawartego pomiędzy tagami strukturalnymi, z niewielką ilością odnośników.
 - Można wyróżnić przynajmniej tytuł notki oraz odnośnik do komentarzy.
 - Odpowiadające tagi strukturalne mają identyczne właściwości - klasy kaskadowych arkuszy stylu.
 - Kolejne struktury umieszczone są w kodzie jedna pod drugą oraz wszystkie mają wspólnego rodzica.
 - Powtarzalność struktury wynosi co najmniej pięć (założenie heurystyczne).
2. Odnalezienie w drzewie układu węzłów, który reprezentuje menu strony (schematyczne odnośniki występujące jeden pod drugim i prowadzące głównie do adresów w domenie lokalnej - archiwum oraz kategorie).
 3. Badanie wzajemnych pozycji odnalezionych kolumn - powinny być umieszczone horyzontalnie względem siebie.
 4. Odnalezienie nagłówka strony, czyli największego elementu graficznego / tekstowego na stronie, występującego pojedynczo.
 5. Badanie wzajemnych pozycji przypuszczalnego nagłówka strony oraz kolumn zawierających menu i wpisy. Nagłówek strony musi znajdować się ponad pozostałymi kolumnami.

Jeżeli wszystkie elementy udało się odnaleźć oraz spełnione są przytoczone założenia przestrzenne, to witryna kategoryzowana jest jako blog.

Strony informacyjne

Strony informacyjne charakteryzują się liczną zawartością odnośników w domenie lokalnej. Podobnie jak w przypadku blogów, dominuje układ dwukolumnowy, przy czym kolumna szersza jest zazwyczaj nieustrukturalizowana, natomiast kolumna węższa zawiera wiele odnośników jeden pod drugim. Logotyp nie jest dominującym elementem strony i ciężko wyznaczyć go na bazie wielkości, zazwyczaj położony jest w lewym górnym rogu witryny. Ponadto, strony informacyjne charakteryzują się wykorzystaniem trzypoziomowej strukturyzacji oferowanej przez HTML5, za pomocą tagów `<div>`, `<article>` oraz `<section>`.

Algorytm rozpoznawania stron informacyjnych będzie działał według następującego schematu:

1. Zbadanie proporcji zwykłego tekstu występującego w ramach odnośników (`<a>...`) oraz innych tagów zawierających tekst widoczny na witrynie.
2. Określenie częstotliwości wykorzystania tagów `<article>` oraz `` do strukturyzacji zagnieżdżonych odnośników do domen lokalnych (newsy).
3. Odnalezienie wąskiej kolumny zawierającej odnośniki do wiadomości.

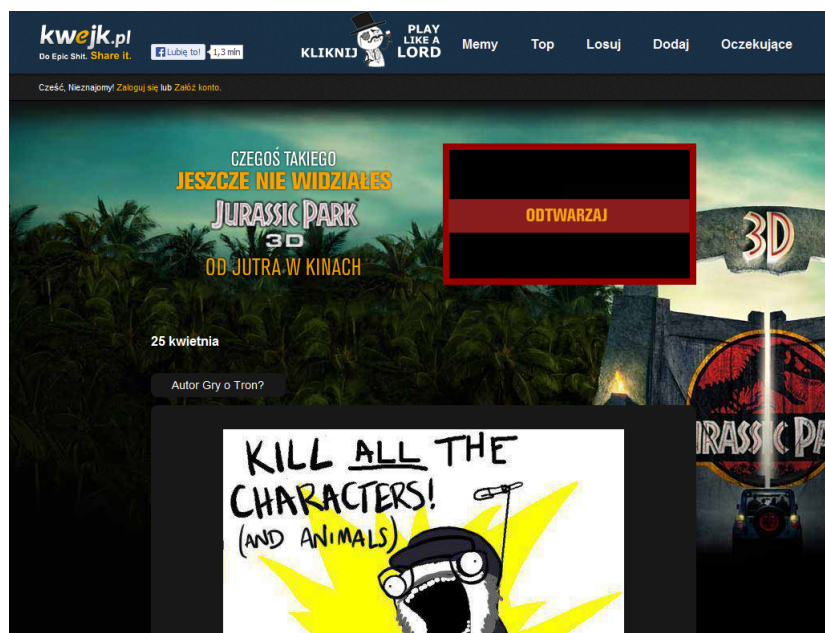
Jeżeli proporcje w punkcie 1. są $\gg 1$ oraz częstotliwość w punkcie 2. jest wysoka, a także udało odnaleźć się kolumnę z punktu 3. to strona klasyfikowana jest jako strona informacyjna.

„Kwejki”

Strony typu *kwejk* charakteryzują się występowaniem pojedynczej, centralnej kolumny, która zawiera dużych rozmiarów obrazy umieszczone jeden pod drugim. Obrazy te są najczęściej prezentowane w ramach jednego „pudełka”, a więc są osadzone w regularnej strukturze tagów HTML. Ponad obrazami znajduje się logotyp oraz menu strony, natomiast pod nimi umieszczone jest małe menu nawigacyjne, prezentujące kolejne liczby (strony). Ogólny szablon stron typu *kwejk* przedstawia rysunek 7, natomiast rysunki 8 oraz 9 pokazują przykłady takich stron.



Rysunek 7: Szablon stron typu *kwejk*.



Rysunek 8: Portal społecznościowy <http://kwejk.pl> oparty na grafikach.

Algorytm kategoryzacji „kwejków” polega na odnalezieniu powtarzającej się struktury węzłów, o wspólnym rodzicu i identycznych właściwościach, w każdym z których osadzona jest pojedyncza grafika. Jeżeli taka struktura występuje, badane jest położenie tej struktury na stronie. W przypadku, gdy kolumna jest wyśrodkowana, strona może zostać zakwalifikowana jako „kwejk”.

Sklepy internetowe

Strony sklepów komputerowych lub ze sprzętem elektronicznym charakteryzują się *kafelkowym* ułożeniem głównej części witryny. Każdy *kafelek* odpowiada specyficznej kategorii towaru, jest ozdobiony pojedynczą grafiką i zawiera wiele odnośników do stron w domenie lokalnej, które odpowiadają podkategoriom produktów. Cechą charakterystyczną jest również zwielokrotniona nawigacja obejmująca:

- klasyczne menu po lewej lub prawej stronie,
- dwa paski nawigacyjne umieszczone bezpośrednio nad i pod *kafelkami*,
- menu z tabularyczną strukturą linków znajdujące się w stopce strony.

Istotnym elementem układu strony jest również duży logotyp sklepu, umieszczony w górnej części strony.

Algorytm klasyfikacji sklepów internetowych oparty będzie na poszukiwaniu *kafelków*. Do ich odnalezienia wykorzystane zostanie poszukiwanie regularnej struktury w drzewie, o następującej charakterystyce:

- identyczne klasy elementów rozmieszczonych w układzie tabelarycznym,



Rysunek 9: Portal społecznościowy <http://demotywatory.pl> oparty na grafikach.

- każdy element zawiera niewielkich rozmiarów grafikę,
- każda kategoria opisana jest promocyjnym towarem - wyszukiwanie wzorca regularnego <kwota> <waluta> w tekście.

Testowanie

Testy opierać się będą na wykorzystaniu gotowego zestawu wstępnie skategoryzowanych witryn. Program testowy, dla każdej strony w zbiorze, wywoła aplikację kategoryzującą i na podstawie danych oceni, czy witryna została przydzielona do prawidłowej grupy. Na tej podstawie zostanie obliczona procentowa skuteczność każdego z algorytmów kategoryzacji. Zbiór witryn testowych dobierany będzie według następujących kryteriów:

- Reprezentatywny (liczący przynajmniej 100 elementów) zbiór próbek dla każdej z kategorii.
- W ramach każdego zestawu występują strony w przynajmniej trzech różnych językach naturalnych.
- Dodatkowy, liczny zbiór witryn, których nie można przydzielić do żadnej z kategorii.

Przykładowe, niepełne listy skategoryzowanych witryn zostały dołączone do dokumentacji.

Dla każdego z algorytmów zostaną wyznaczone następujące zbiory:

- TP (true positives) - poprawne przydzielenie witryny do kategorii (algorytm wskazał, że witryna należy do kategorii, gdy w rzeczywistości do niej należy),
- TN (true negatives) - poprawne nieprzydzielenie witryny do kategorii (algorytm wskazał, że witryna nie należy do kategorii, gdy w rzeczywistości do niej nie należy),
- FP (false positives) - błędne przydzielenie witryny do kategorii (algorytm wskazał, że witryna należy do kategorii, gdy w rzeczywistości do niej nie należy),
- FN (false negatives) - błędne nieprzydzielenie witryny do kategorii (algorytm wskazał, że witryna nie należy do żadnej kategorii, gdy w rzeczywistości do niej należy).

Na podstawie tych wartości, dla każdego algorytmu zostaną wyznaczone:

- precyzja - $\frac{|TP|}{|TP|+|FP|}$,
- zupełność - $\frac{|TP|}{|TP|+|FN|}$,
- dokładność - $\frac{|TP|+|TN|}{|TP|+|TN|+|FP|+|FN|}$,
- zaszumienie - $\frac{|FP|}{|FP|+|TN|}$.